Anything you want as long as you say Petabyte/Petascale

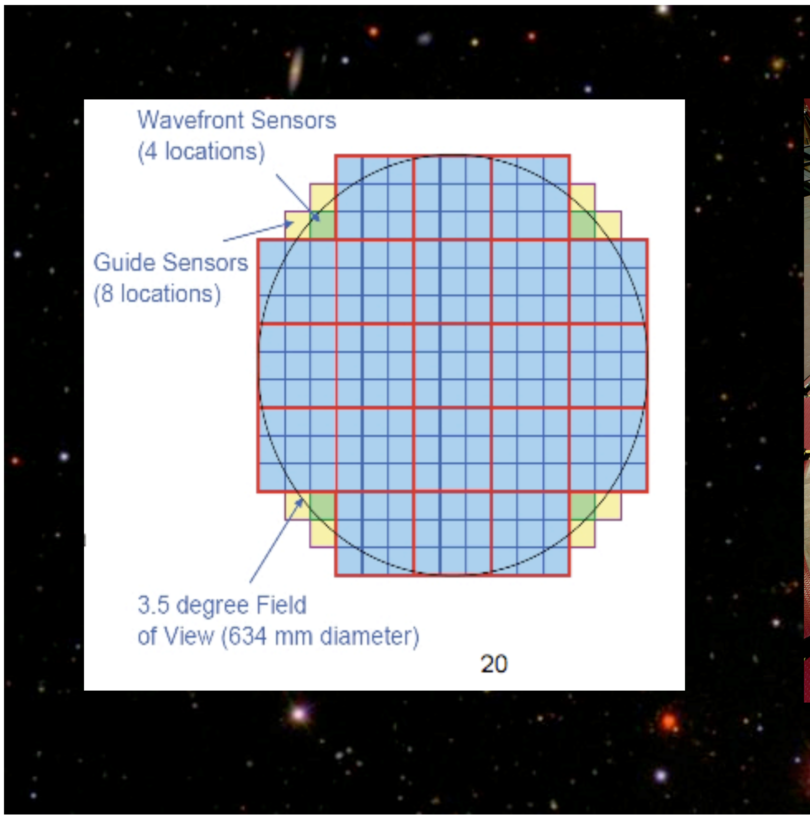
~~Data~~^~~Management and Mining for Ultra-Large Photometric Surveys~~

Andrew Connolly

Department of Astronomy

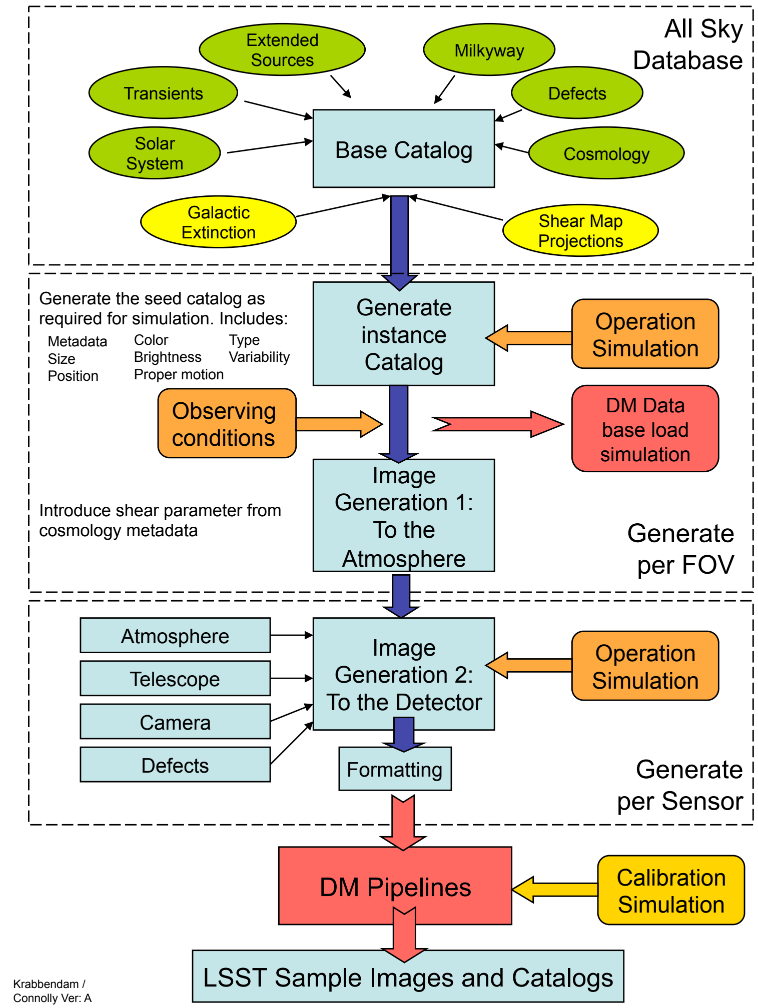
With thanks to: Zeljko Ivezic, John Peterson, Garrett Jernigan, Jim Pizagno, Andy Becker, Andy Rasmussen, Kirk Gilmore, Simon Krughoff, Lynne Jones, Francesco Pierfederici, Phil Pinto, Alan Meert

Simulating the Sky



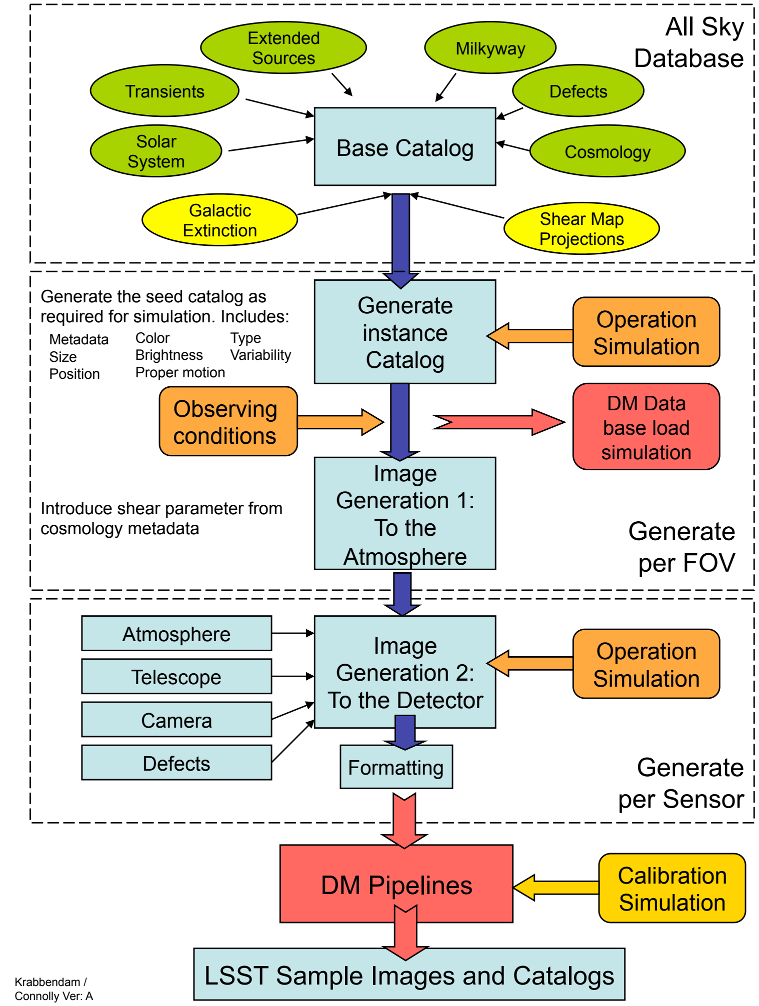
Simulating a Petabyte Data Stream

- **LSST data flow**
  - ½ the sky every 3 nights
  - 40 TB of imaging per night
  - 10^9 sources a night
  - 10^3 "events"
  - 1000x in 10 years
  - 5 months to watch 1 year of data on an HDTV.
- **Simulation flow with LSST**
  - 1 Petabyte after year one
  - 60 Petabytes of images after 10 years
  - Galaxies, stars, weak lensing, extinction, solar system objects… images



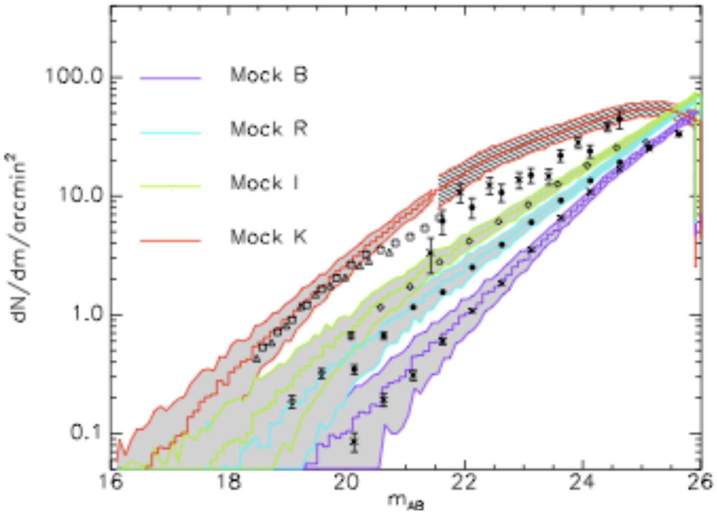
Base Catalog Design

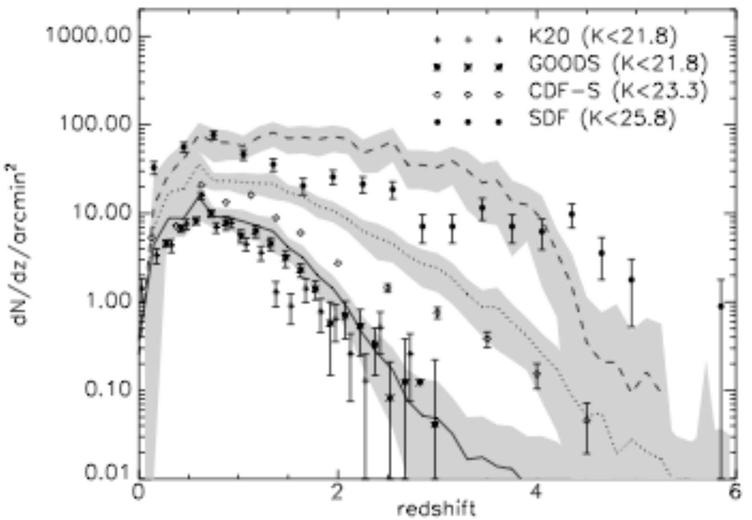
- **Atomic representation of input catalogs**
  - **Evolving Base catalogs**
    - Input cosmologies
    - Milky-way model
    - Extinction screen
    - Shear maps
    - Consistent API
  - **Extended implementation**
    - Defects
    - Moving sources
    - Extended sources (add your own image)
  - **Initial database access**
    - Mysql moving to sqlserver



Cosmology

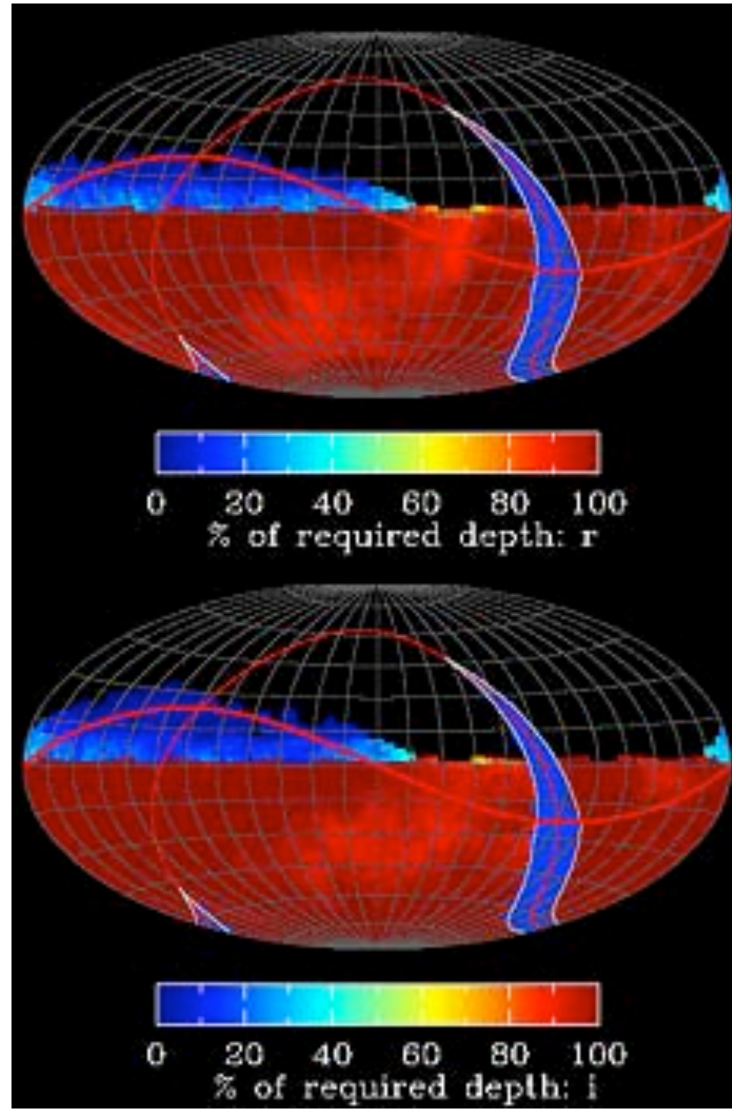
- **Millennium Simulations**
  - **Kitzbichler and White (2006)**
    - 6 fields, 1.4x1.4 deg per field
    - $6x10^6$ source per catalog
    - Based on Croton et al (2006) and De Lucia and Blaizot (2006) models
    - r<26 magnitude limit
    - z>4 redshift limit
    - BVRIK Johnson and griz SDSS
    - Extended to fit LSST u,g,r,i,z,y3
    - Derived SED for all sources





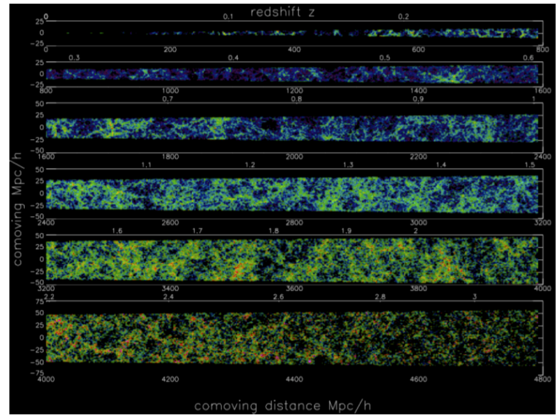
"Observing" the LSST Simulation

- An instance catalog
  - Sampling the base catalog using outputs from OpSim
  - Position, atmosphere, filter, time
  - Sample light curve as well as static populations
  - r<28 to simulate the sky accurately
  - Output catalogs and metadata for the photon tracing simulations
  - SQL and python interfaces



From Catalogs to Photons and Back

- **Ray Tracing the sky**
  - **High fidelity simulator**
    - **Based on Physics of atmosphere, telescope, camera, detector**
    - **Input catalog and images with associated SEDs**
    - **Produce realistic images**
    - **Understand characteristics of the PSF**
    - **Model thermal effects**
    - **Wavelength dependent effects**



Catalog Generation (Millennium Simulations)

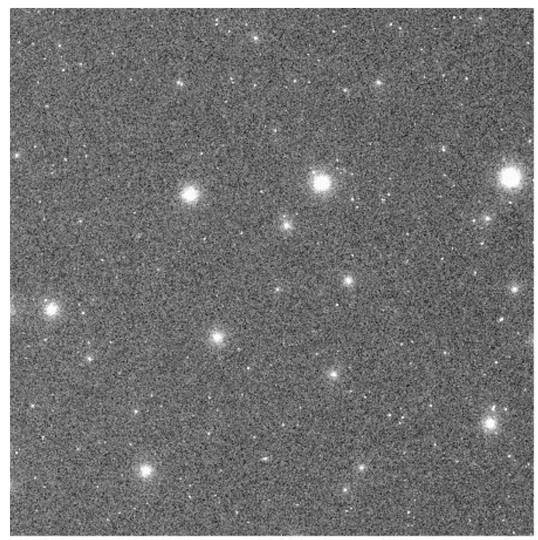
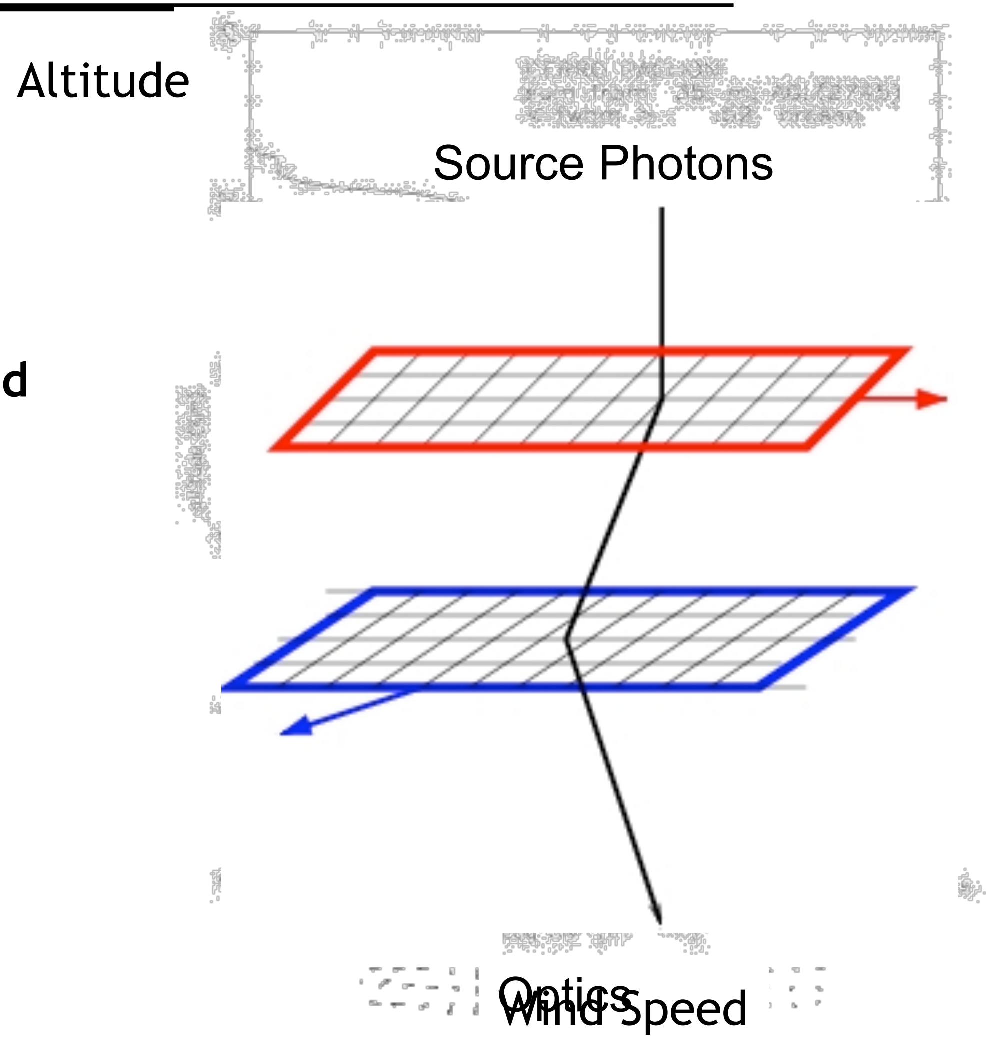


Image Generation (Full Photon Ray Tracing)

Atmosphere Models & Kolmogorov turbulence

- **Turbulent screens**
  - **Data from Cerro Pachon**
  - **Arbitrary number of screens**
  - **Arbitrary velocity vectors**
  - **Photons ray traced and shifted**
  - **Vector Screen:**
    - 2048 squared
    - 0.1m/pixel

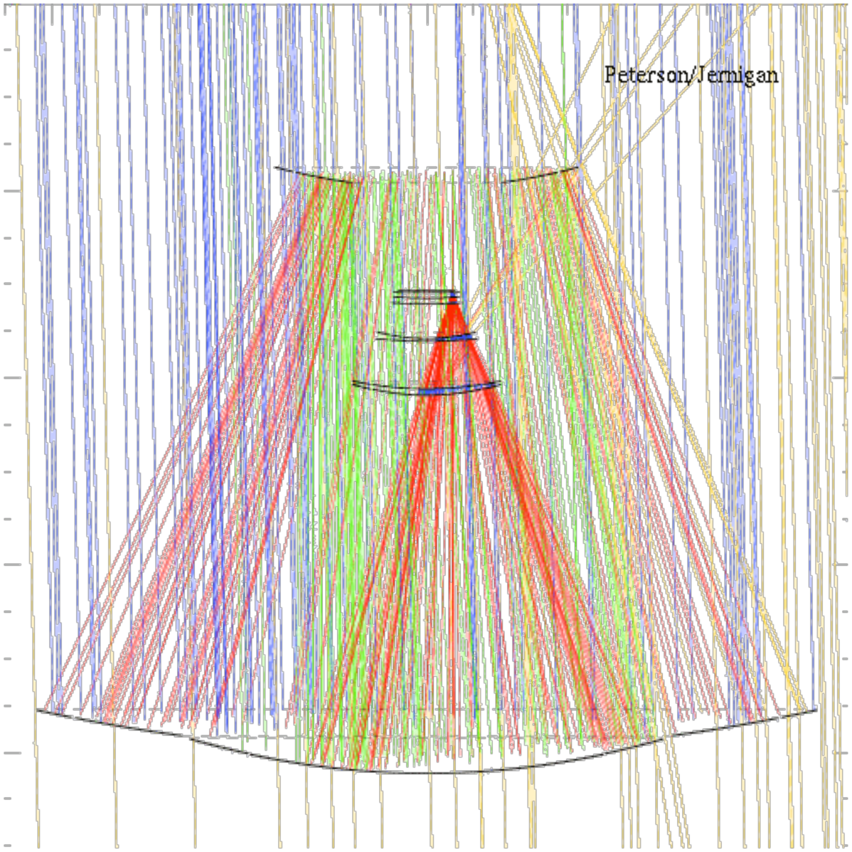


Vernin et al., Gemini RPT-A0-G0094

Jernigan, Peterson and Rasmussen

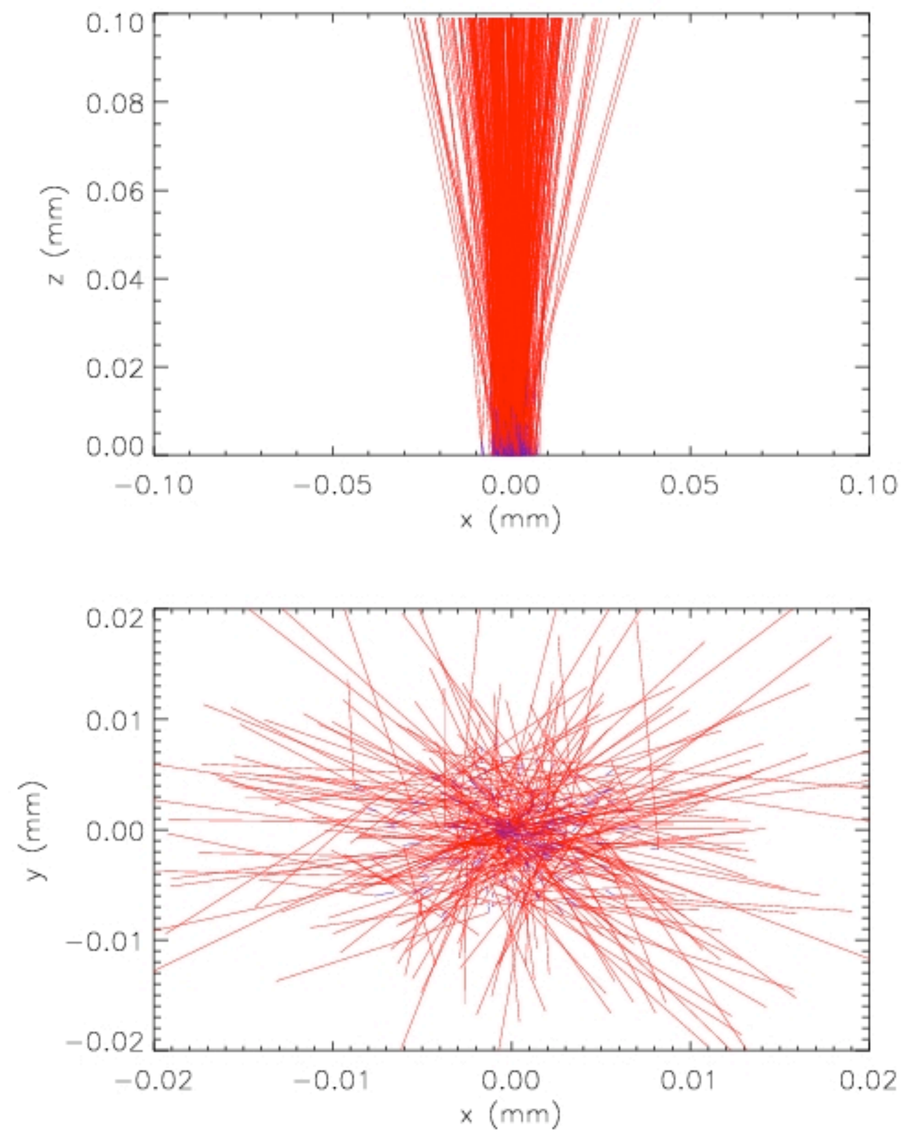
Telescope Optics

- **Telescope model**
  - LSST baseline design
  - Input zemax model
  - Fast ray trace
  - Calculates ray intercepts
  - Fast reflection and refraction algorithms
  - Wavelength-dependent effects

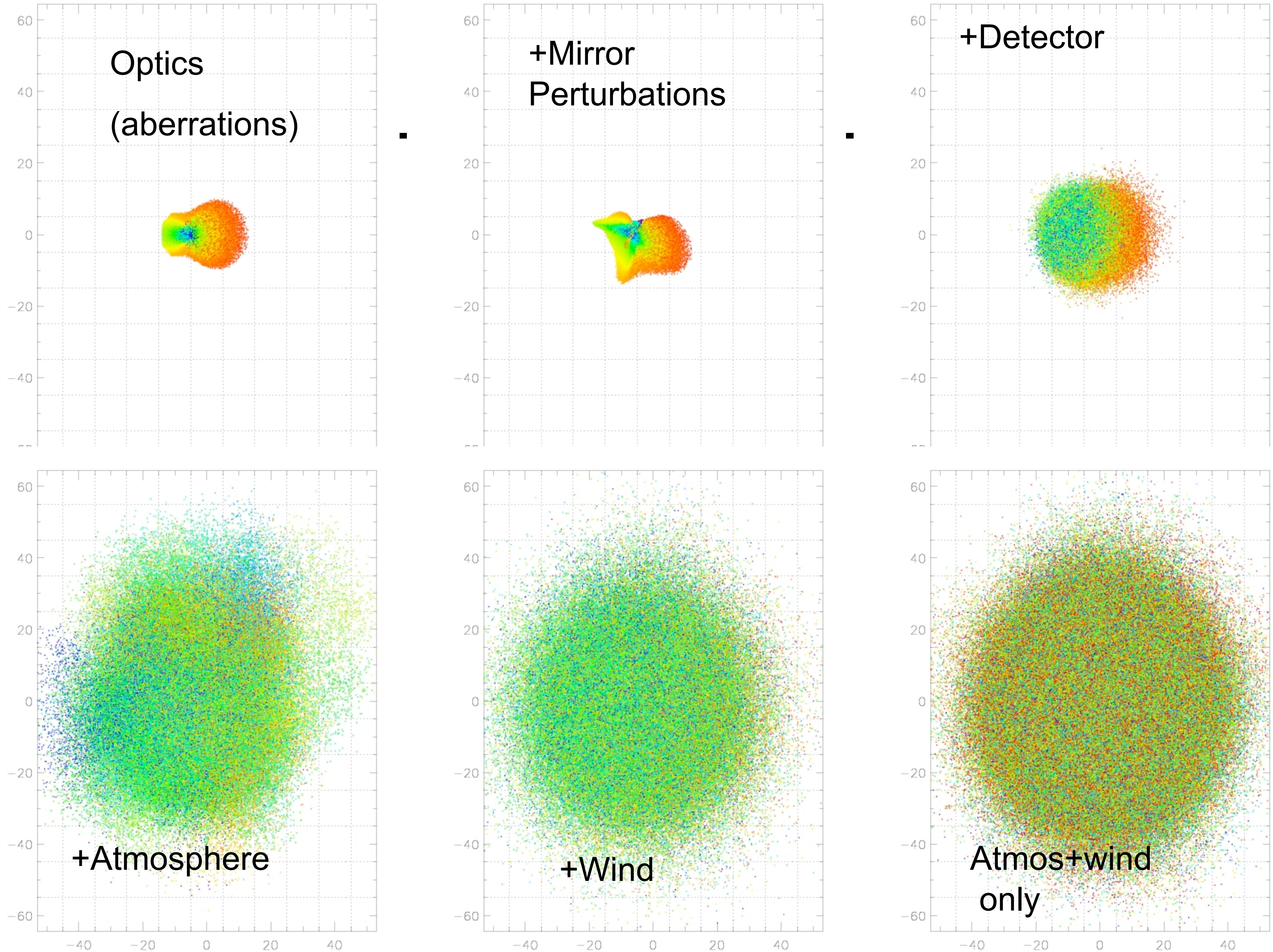


Camera and Detector model

- **Focal plane model**
  - **Modeling for 200 CCDs in focal plane**
  - **Incorporates chip gaps, boron implants**
  - **Chip pistoning and surface effects**
- **Detector model**
  - **Refraction for light entering the Si surface**
  - **Photon interaction (wavelength and temperature dependence)**
  - **Lateral charge diffusion**



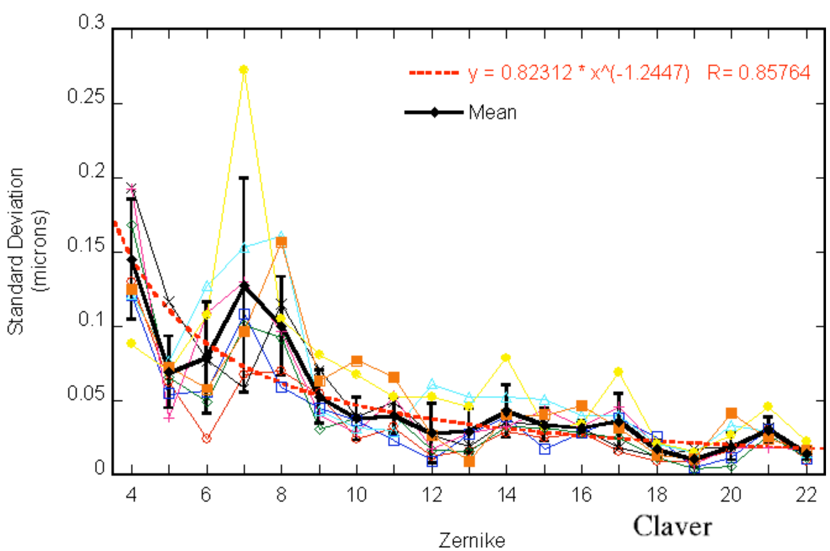
Rasmussen and Gilmore



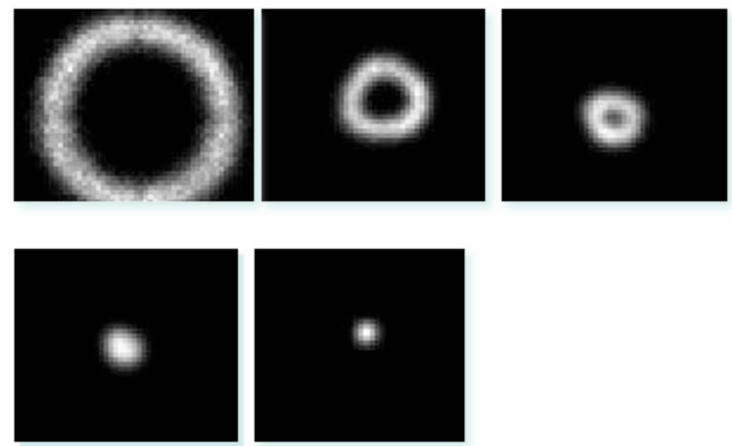
Optics
(aberrations)
+Mirror
Perturbations
+Detector
+Atmosphere
+Wind
Atmos+wind
only

Examples: Thermal and mechanical distortions

- **Simulating perturbations**
  - **Each optic has 6 dof (decenter, defocus, three euler angles)**
    - Perturbations are placed on the three mirrors using a Zernike expansion to simulate the possible residual control system errors each mirror can have an arbitrary amplitude code goes up to 5th order polynomials



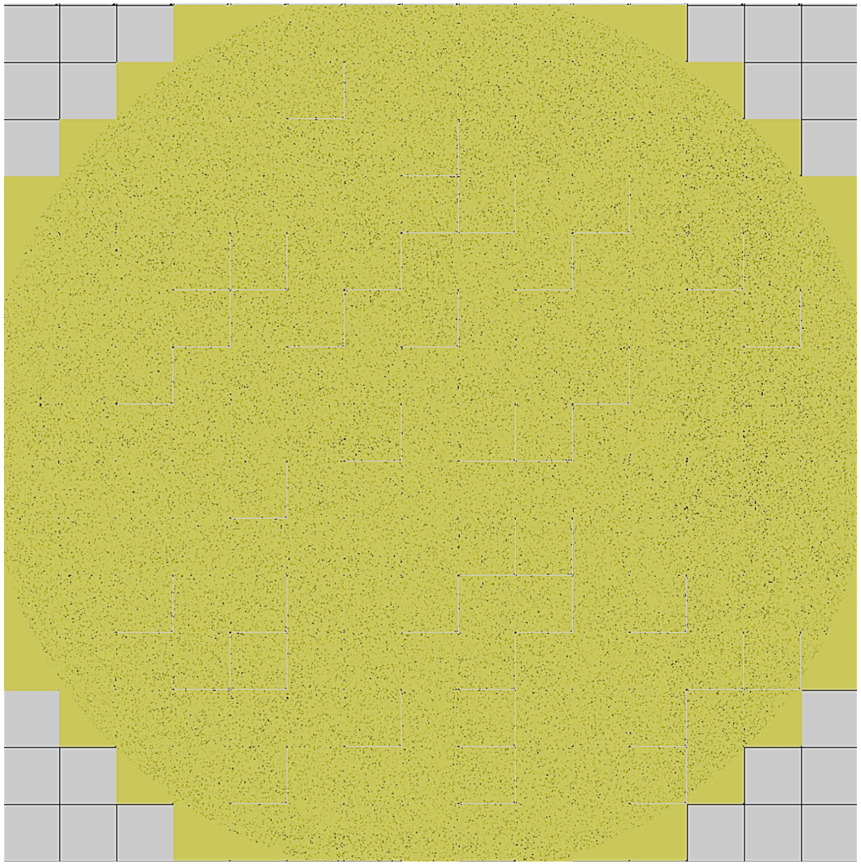
Perturbation spectrum
From Claver



e.g. Mirror Defocus

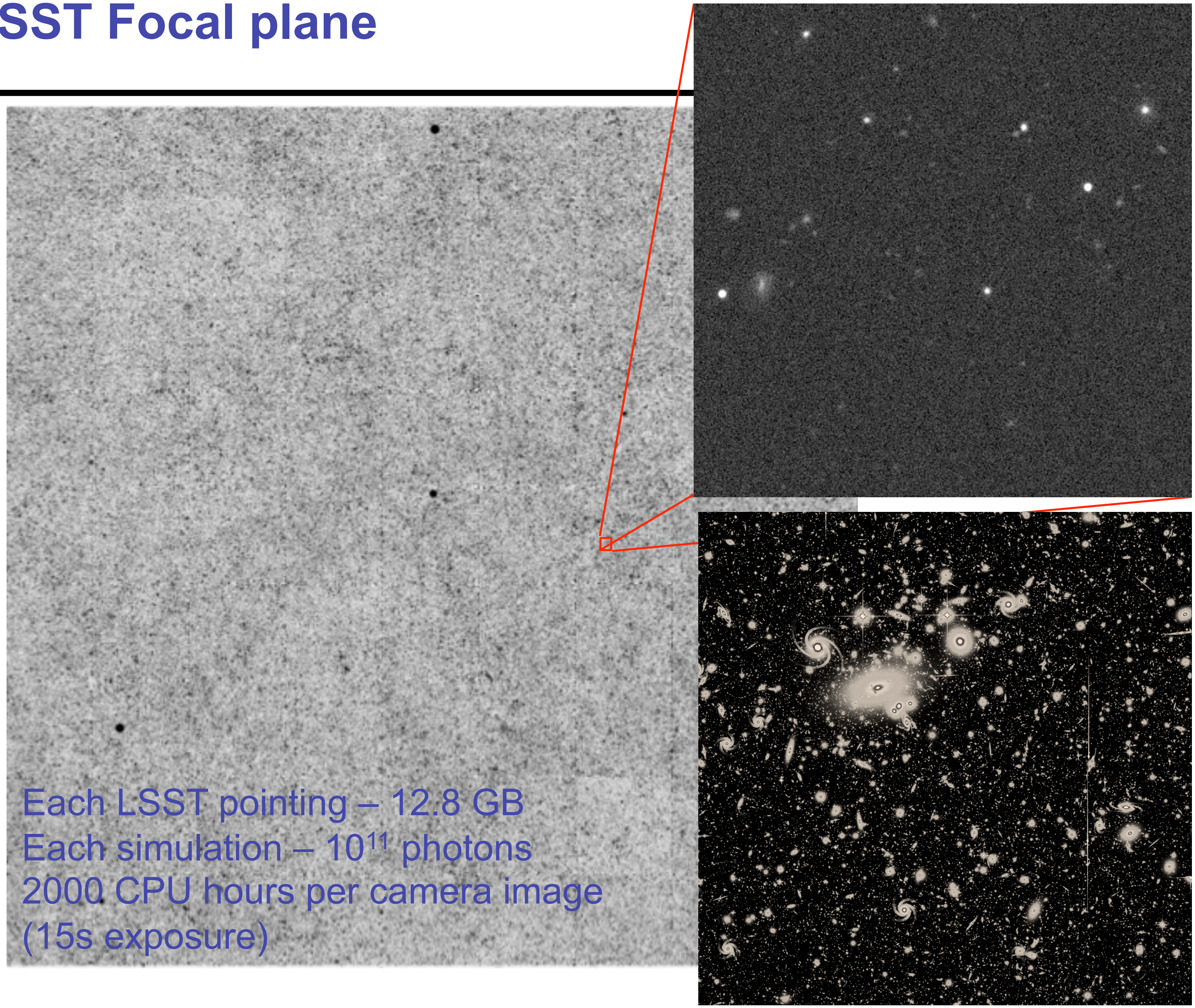
LSST focal plane

- **Simulating 3.2 Gpixels**
  - **10^{11} photons per focal plane**
  - **12.8 GB per image**
  - **2000 CPU hrs per focal plane**
  - **Simulated CCD at a time**
    - Moving to amplifier granularity
- **Distributing the load**
  - **Condor pool as the initial pipeline (Purdue)**
  - **Refactored to run under Hadoop (Mapreduce)**
    - Finer granularity
  - **Run across 1000 cores and up**



Meert (Purdue)

An LSST Focal plane



Each LSST pointing – 12.8 GB
Each simulation – 10^{11} photons
2000 CPU hours per camera image
(15s exposure)

Challenges ahead (lessons we will learn)

- **Supporting a full end-to-end simulation**
  - **Database access for science collaborations**
    - Derived catalogs and images with variability
    - Enabling science with LSST ahead of time
    - Many different use cases
  - **Challenge of a fully distributed system**
    - Data size is a challenge – simulating LSST database 7 years ahead of time
    - CPU load is a challenge – looking at map-reduce, Dryad, Hadoop as a distributed system
    - Compute and storage resources

Give me your tired, your poor,
Your free cycles and particles yearning to breathe free,
The wretched refuse of your teeming shore.

Emma Lazarus (sort of)